The Role of Representation Information in the Digital Curation of Crystallography Data

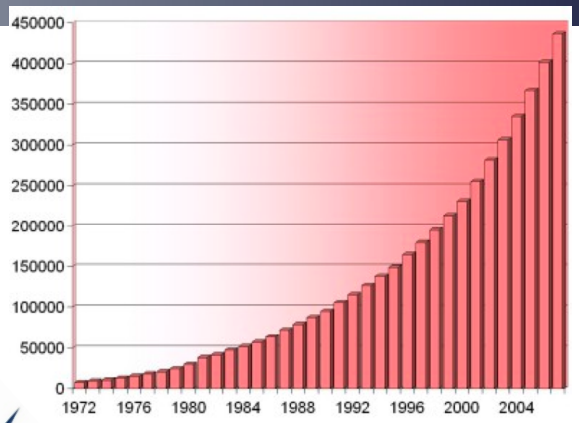
Manjula Patel[1], Simon Coles[2], David Giaretta[3], Stephen Rankin[3], Brian McIlwrath[3]

[1]UKOLN, University of Bath, UK
[2]EPSRC NCS, University of Southampton, UK
[3]Science & Technology Facilities Council, UK

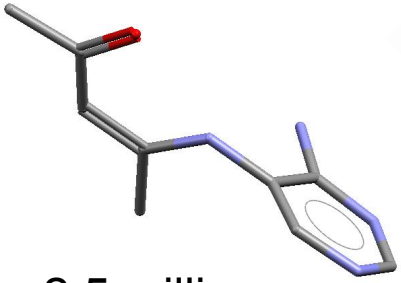
5th IEEE International Conference on e-Science
9-11th December 2009
Oxford, UK

Data Deluge

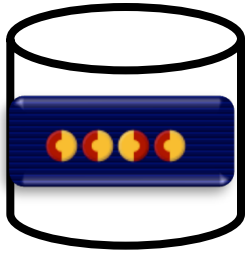
40 years ago a PhD student would determine about 3 crystal structures for their thesis – this can now be easily achieved in a day



35 million



2.5 million



0.5 million

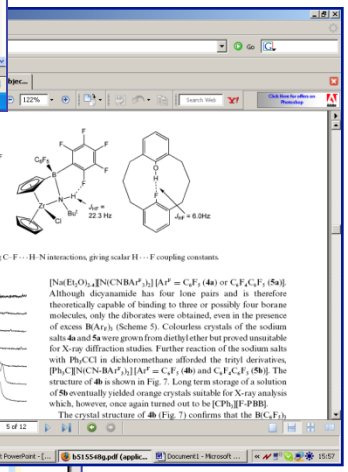
'Few thousand'

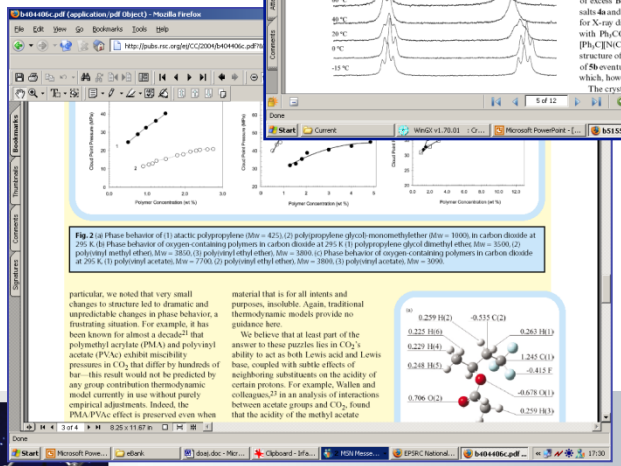
The primary cause is the current data publication process, which is tied to journal articles and peer review

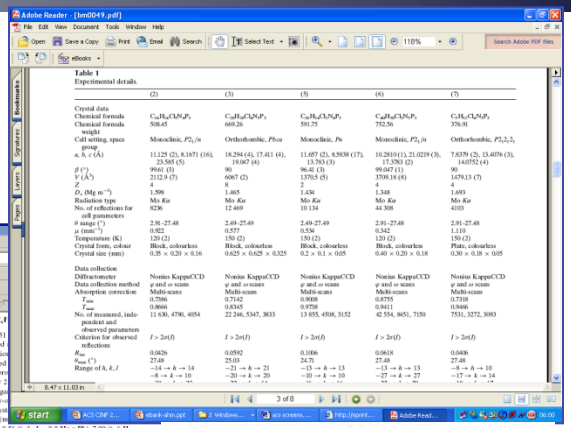
The Solution

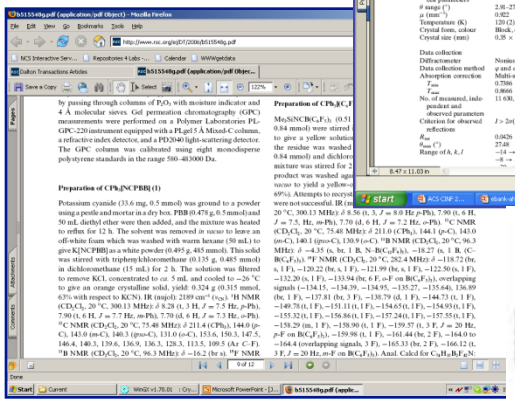
Intellect & Interpretation (Journal article, report, etc)

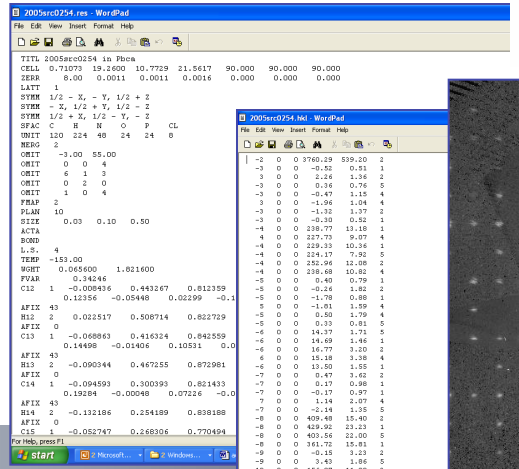
Underlying data (Institutional data repository)

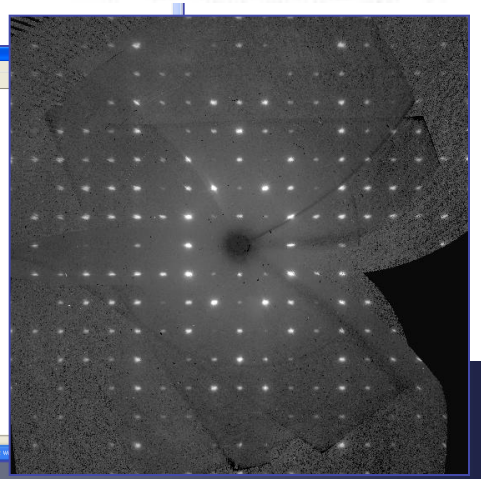












eCrystals Federation

5th IEEE eScience Conference, 9-11th December 2009, Oxford, UK

The eCrystals Data Repository

eCrystals — University of Southampton

Home | About | Browse by Year | Browse by People

Logged in as Dr Richard A Stephenson | Manage deposits | Profile | Saved searches | Review | Admin | Logout | Search

2,2-trimethylenedioxy-4,4,6,6-tetrachlorocyclotriphosphazene

Sample Originator: A. Kilic[a], M. Odlyha[b], A. Uslu[a], David B. Davies[b] and R.A. Shaw[b].

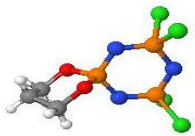
Data Collection: Mark E. Light[c], Simon J. Coles[c] and Susanne. L. Huth[c]

Structure Determination: Simon J. Coles[c], Michael B. Hursthouse and J.S. Rutherford.

Gebze Institute of Technology[a]
Birkbeck College[b]
University of Southampton[c]

$C_3H_6Cl_4N_3O_2P_3$

InChI=1/C3H12Cl4N3O2P3/c4-13(5)8-14(6,7)10-15(9-13)11-2-1-3-12-15/h8-10,13-15H,1-3H2



Jmol

Identification Number:	10.3737/ecrystals.chem.soton.ac.uk/300
Controlled Keywords:	cyclophosphazene, phase transition, variable temperature
Date Created:	28 March 2007
Deposited On:	*21 Apr 2008 15:56*
Deposited By:	Dr Simon J Coles

Data collection parameters

Chemical formula	C3 H6 Cl4 N3 O2 P3
Crystallisation Solvent	
Crystal morphology	Rod
Crystal system	Orthorhombic
Space group symbol	Pna2(1)
Cell length a	13.4804(14)
Cell length b	10.6442(9)
Cell length c	8.8479(7)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	274(2)

Refinement results

Solution figure of merit	0.0569
R Factor (Obs)	0.0334
R Factor (All)	0.0380
Weighted R Factor (Obs)	0.0871
Weighted R Factor (All)	0.0905

Citation: Kilic, A. and Odlyha, M. and Uslu, A. and Davies, David B. and Shaw, R.A. and Light, Mark E. and Coles, Simon J. and Huth, Susanne. L. and Hursthouse, Michael B. and Rutherford, J.S. (2007) University of Southampton, Crystal Structure Report Archive. (doi:10.3737/ecrystals.chem.soton.ac.uk/300)
Export as: EndNote BibTeX ASCII Citation

Available Files

Final Result

2005sjc0007.cif	11k
2005sjc0007.cml	4k
2005sjc0007.fcf	138k

Validation

2005sjc0007_checkcif.htm	9k

Refinement

2005sjc0007.res	5k
2005sjc0007_xl.lst	29k

Solution

2005sjc0007.prp	5k
2005sjc0007_xs.lst	44k

Processing

2005sjc0007.hkl	532k
2005sjc0007.htm	11k
2005sjc0007_0kl.jpg	91k
2005sjc0007_h0l.jpg	87k
2005sjc0007_hk0.jpg	79k

Data Collection

2005sjc0007_crystal.jpg	17k

Other Files

2005sjc0007.doc	186k
2005sjc0007.inchi	1k
2005sjc0007.ins	4k
2005sjc0007.mol	2k
2005sjc0007.p4p	1k
2005sjc0007_ellipsoid.gif	

Repository Staff Only: item control page

- Quick & simple to deposit
- Software tools
- Laboratory archive
- Community involvement
- ‘Embargo’ facility
- Structured foundations
- Discoverable & harvestable
- Number of file formats

UKOLN

DCC

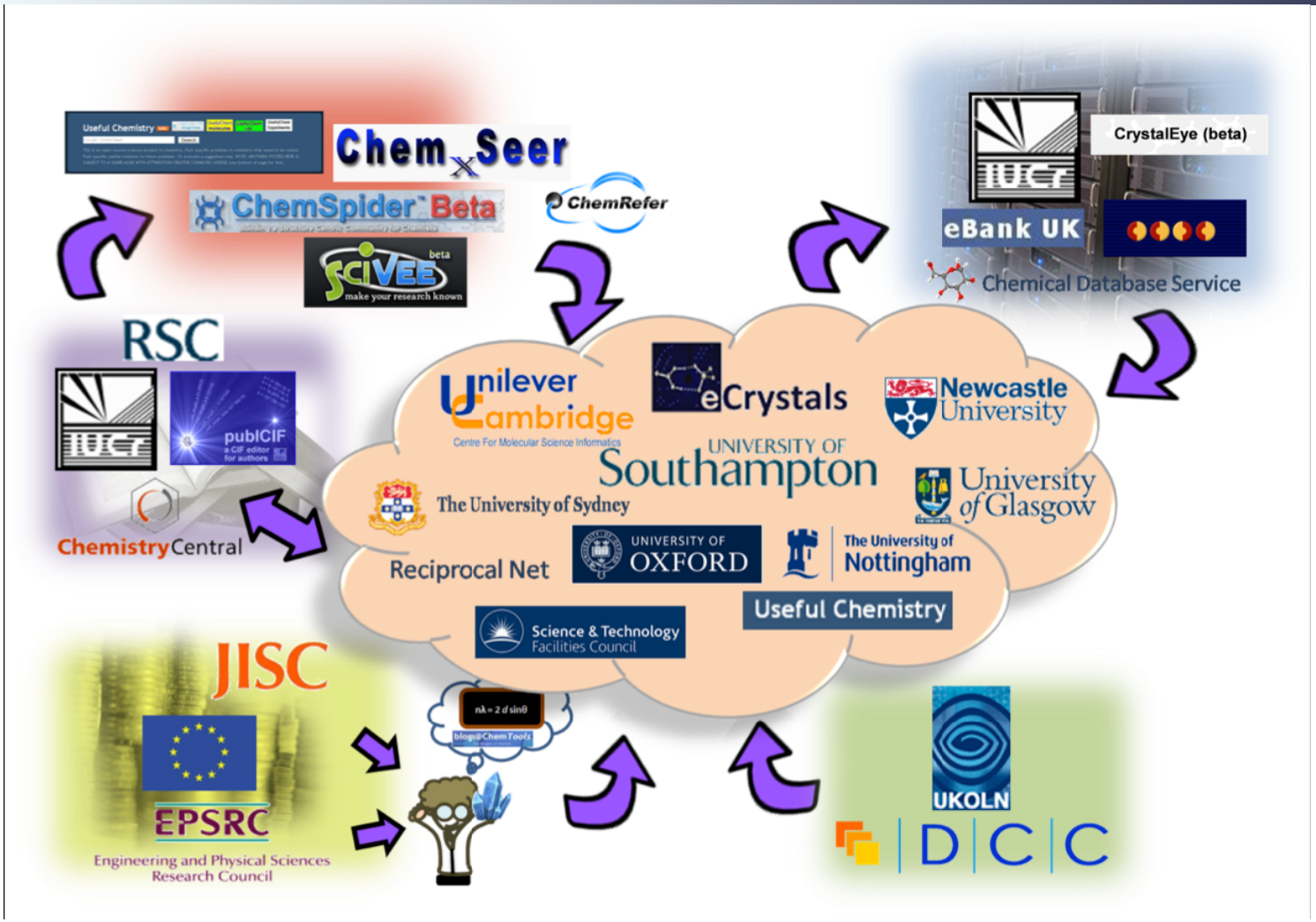
eprints

http://ecrystals.chem.soton.ac.uk

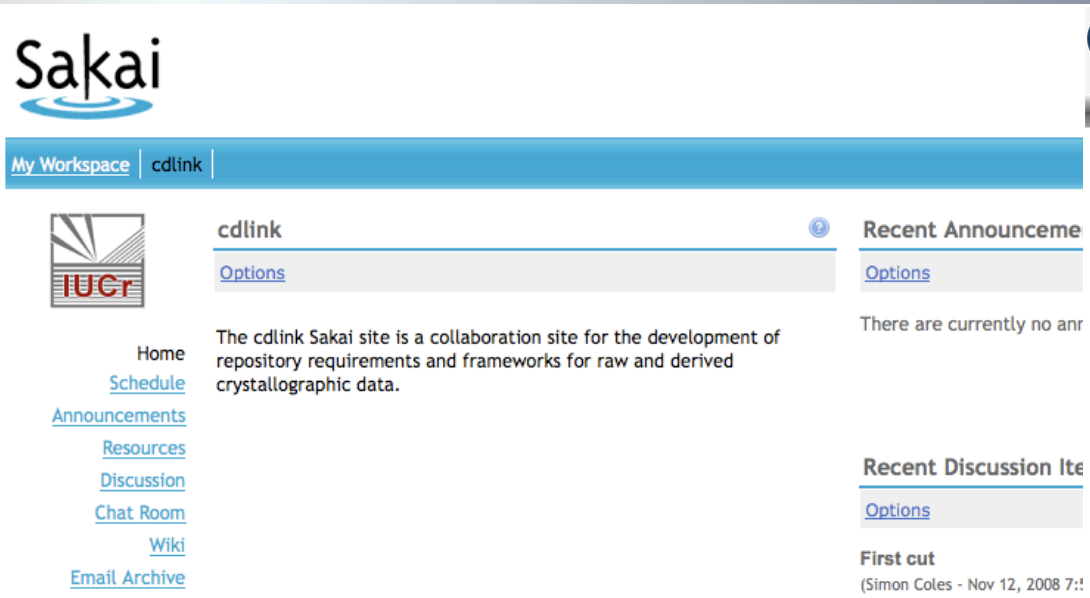
A Thorough Approach to Dissemination

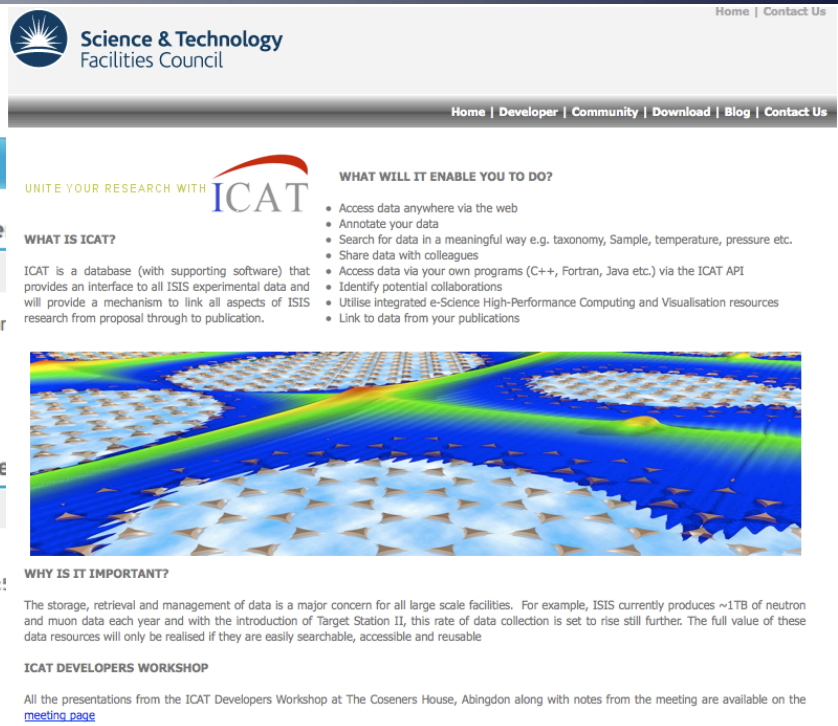
- Using simple Dublin Core protocol (OAI-PMH)
  - Crystal structure
  - Title (Systematic IUPAC Name)
  - Authors
  - Affiliation
  - Creation Date
- Additional chemical information through Qualified Dublin Core
  - Empirical formula
  - International Chemical Identifier (InChI)
  - Compound Class & Keywords
- Specifies which 'datasets' are present in an entry
- Application Profile *http://www.ukoln.ac.uk/projects/ebank-uk/schemas/*
- DOI links *http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145*
- Rights & Citation *http://ecrystals.chem.soton.ac.uk/rights.html*

Building a Federation of Repositories



Crystallography Data Commons





- Generic publication & dissemination
- Domain & context
- Management
- Preservation Data for Crystallography Data
- Core Scientific Data Model

OAIS Background

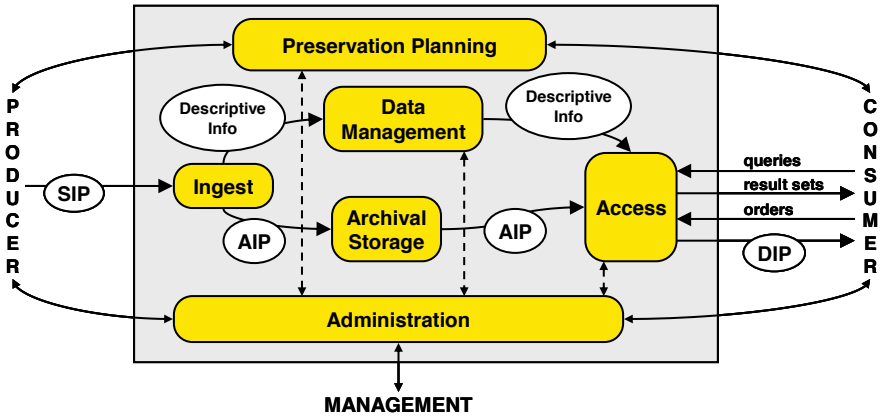
- OAIS -*Reference Model for an Open Archival Information System* http://www.ccsds.org/documents/650x0b1.pdf
- Development led by the Consultative Committee for Space Data Systems (CCSDS)
- Adopted as ISO 14721:2003
- “Open” refers to development of the model in an open forum
- Reference Model, not a blueprint for implementation
- Establishes a common framework of terms and concepts
- Identifies the basic functions of an OAIS
- Defines an information model
- Three major areas of influence:
  - Preservation metadata schemas
  - Architecture and system design
  - Conformance criteria for archival repositories

OAIS Definition and Selected Concepts

- OAIS: "An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community"
- Designated Community: Community of stakeholders and users that the OAIS serves
- Knowledge Base: A set of information, incorporated by a user or system, that allows that user or system to understand the received information
- Information Object: Data Object + Representation Information
- Representation Information: any information required to render, interpret, use and understand digital data
- Information Package: Content Information + Preservation Description Information + Packaging Information (Submission, Archival and Dissemination Information Packages)
- Preservation Description Information: Provenance, Context, Reference, Fixity information

OAIS Functional Entities

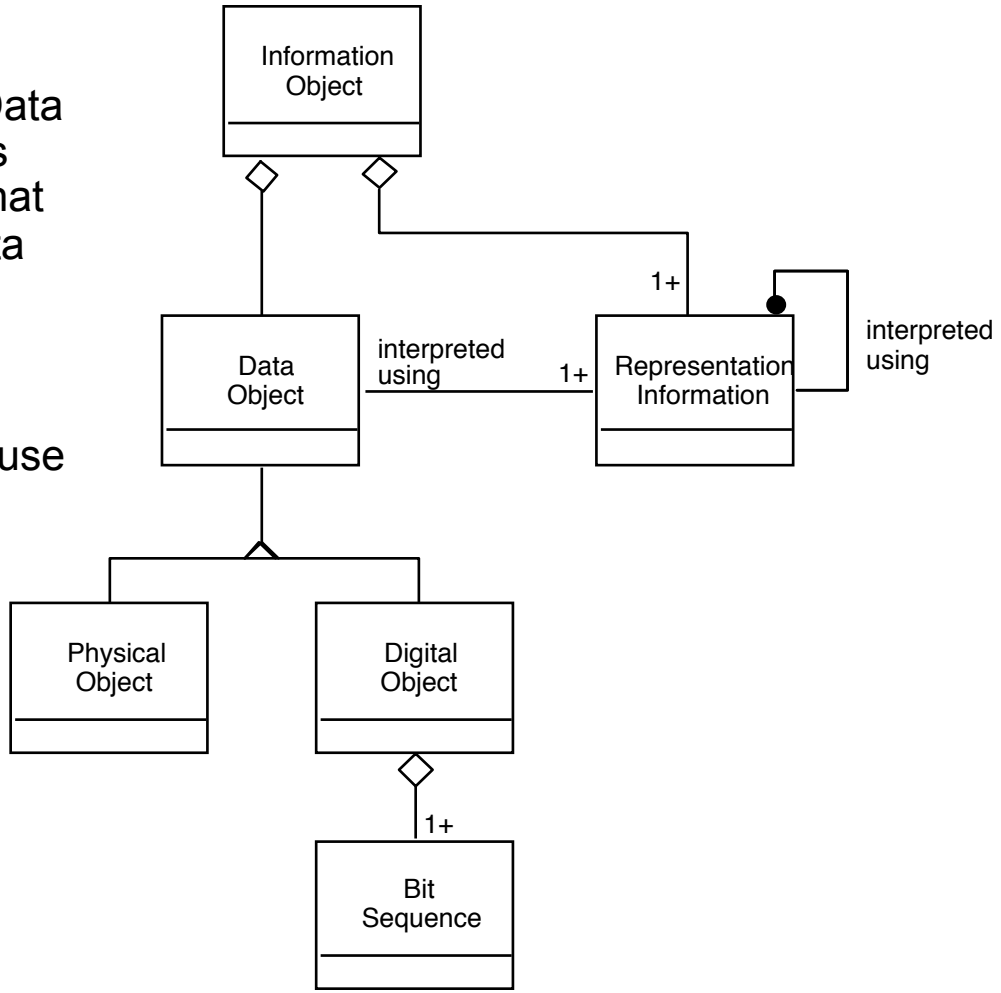
- Ingest: services and functions that accept SIPs from Producers; prepares AIPs for storage, and ensures that AIPs and their supporting Descriptive Information become established within the OAIS
- Archival Storage: services and functions used for the storage and retrieval of AIPs
- Data Management: services and functions for populating, maintaining, and accessing a wide variety of information
- Administration: services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis



- Preservation Planning: services and functions for monitoring the OAIS environment and ensuring that content remains accessible to the Designated Community
- Access: services and functions which make the archival information holdings and related services visible to Consumers

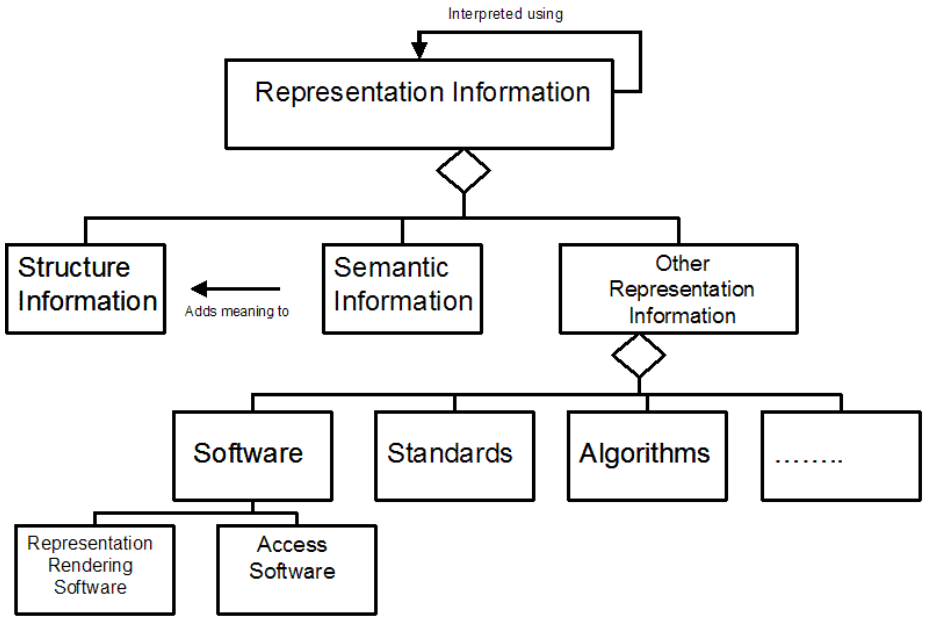
OAIS Information Model

- Information Object is composed of a Data Object that is either physical or digital, as well as the Representation Information that allows for the full interpretation of the data into meaningful information
- Representation Information is *any* information required to render, interpret, use and understand data



OAIS Representation Information (RI)

- Types of RI
  - Structure e.g. file formats for text, images, audio, moving images, datasets, 3D models
  - Semantic e.g. data dictionaries and knowledge organisation systems such as schemata, ontology, metadata vocabularies and thesauri
  - Other e.g. software, algorithms, standards, time dependent information, actions, processes
- RI is recursive in nature; using one element of RI in a meaningful manner may well require further RI, resulting in a RI Network
  - Recursion is terminated based on the designated community's knowledge base
  - Essential that RI itself is curated and preserved to maintain access to data



Registry/Repository of RI (RRoRI)

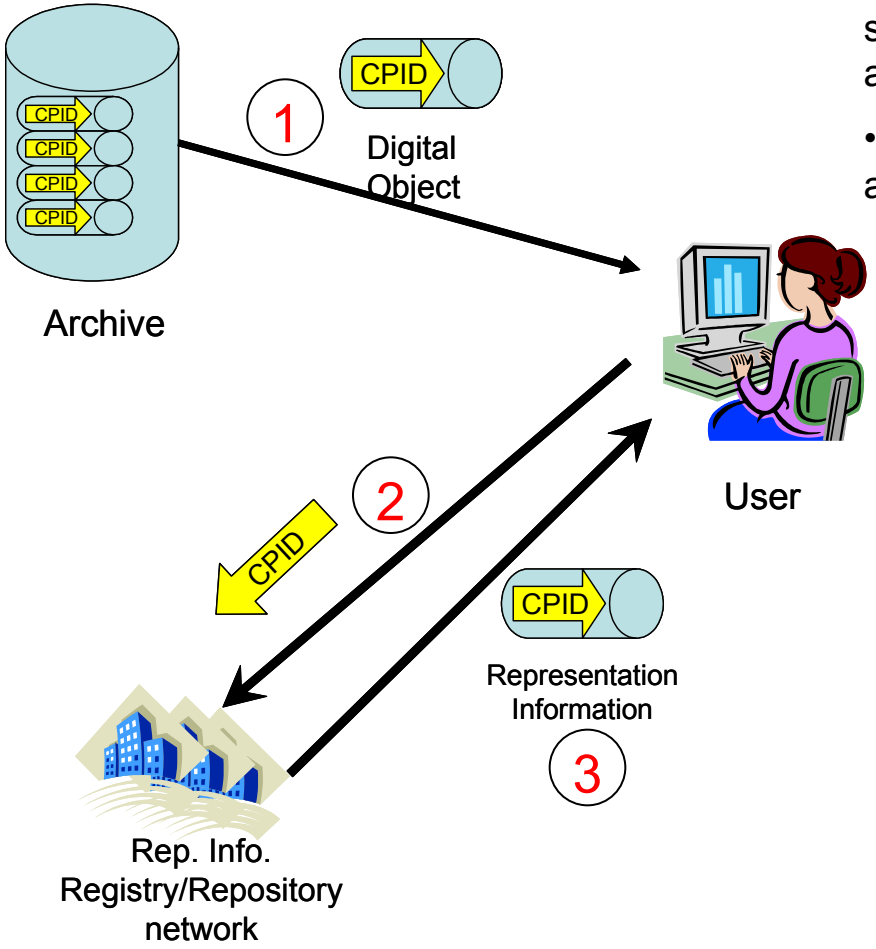
- Development started under the DCC-Development team
- Work now being undertaken jointly with the CASPAR Project
  - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (Integrated Project co-funded by EU FP6 Programme, April 2006)
- Representation Information is the key to long-term access
- RRoRI should itself be a trustworthy OAIS
- Repository: some RI is stored; Registry: links to external RI
- Emphasis on interoperability and automated use
- Vision is to have a global, distributed network of RI
- Provide an infrastructure of reliable and trusted RI for third party use

RRoRI: RI Label & CPID

- Idea of RI is the key
  - Information Object: a specific object to be archived/preserved/ curated
  - RI: all information required to render, interpret, use and understand the object
  - RI Label: used to connect RI to an Information Object
- RI Label serves as a mechanism for accessing RI in RRoRI
  - Label is used to identify relevant RI
  - Provides mechanism for recording individual RI components
- RI Label has a Curation Persistent Identifier (CPID)
  - Used to connect the digital object to the RI Label

Use of CPID

1 User gets data from archive. Data has associated Curation Persistent Identifier (CPID)



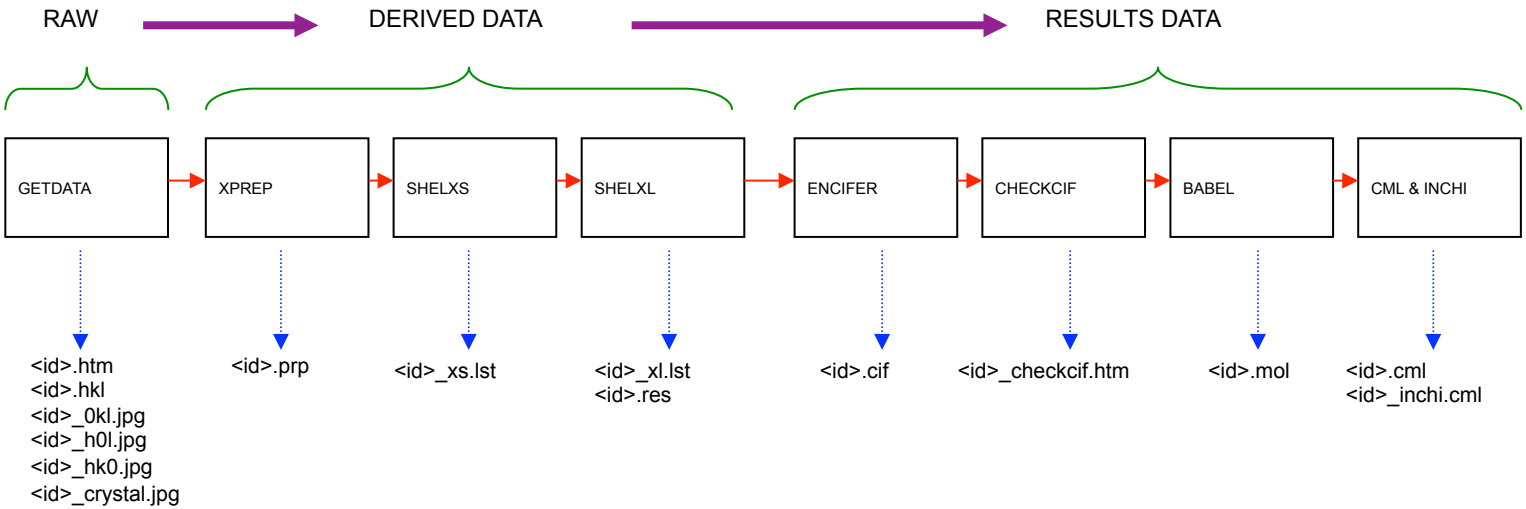
2 User unfamiliar with data so requests RI using CPID

- The Digital Object could have some RI packed with it, as well as a CPID
- CPID supports automated access & processing

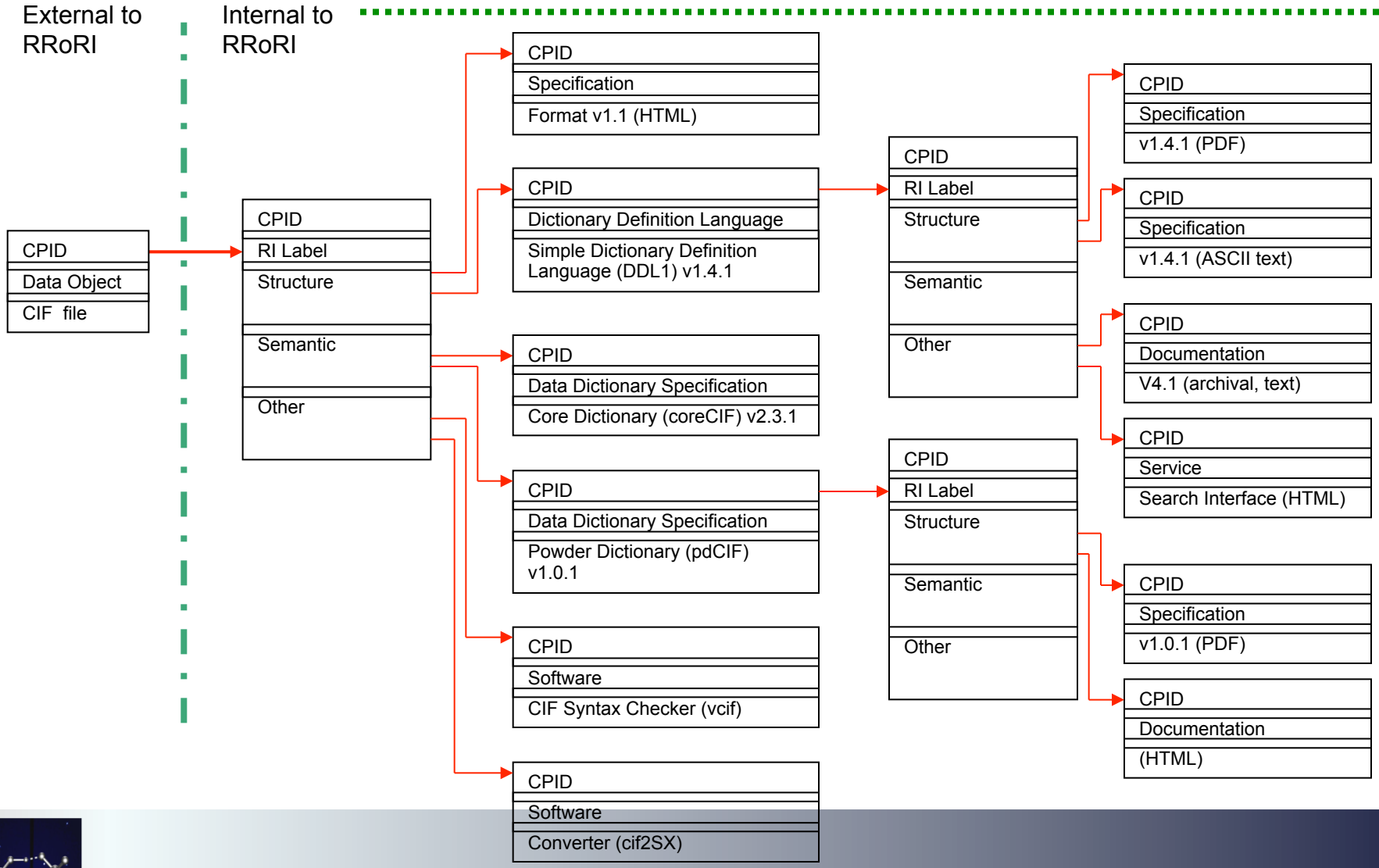
3 User receives RI – which has its own CPID in case it is not immediately usable

Capturing RI: Crystallography Data

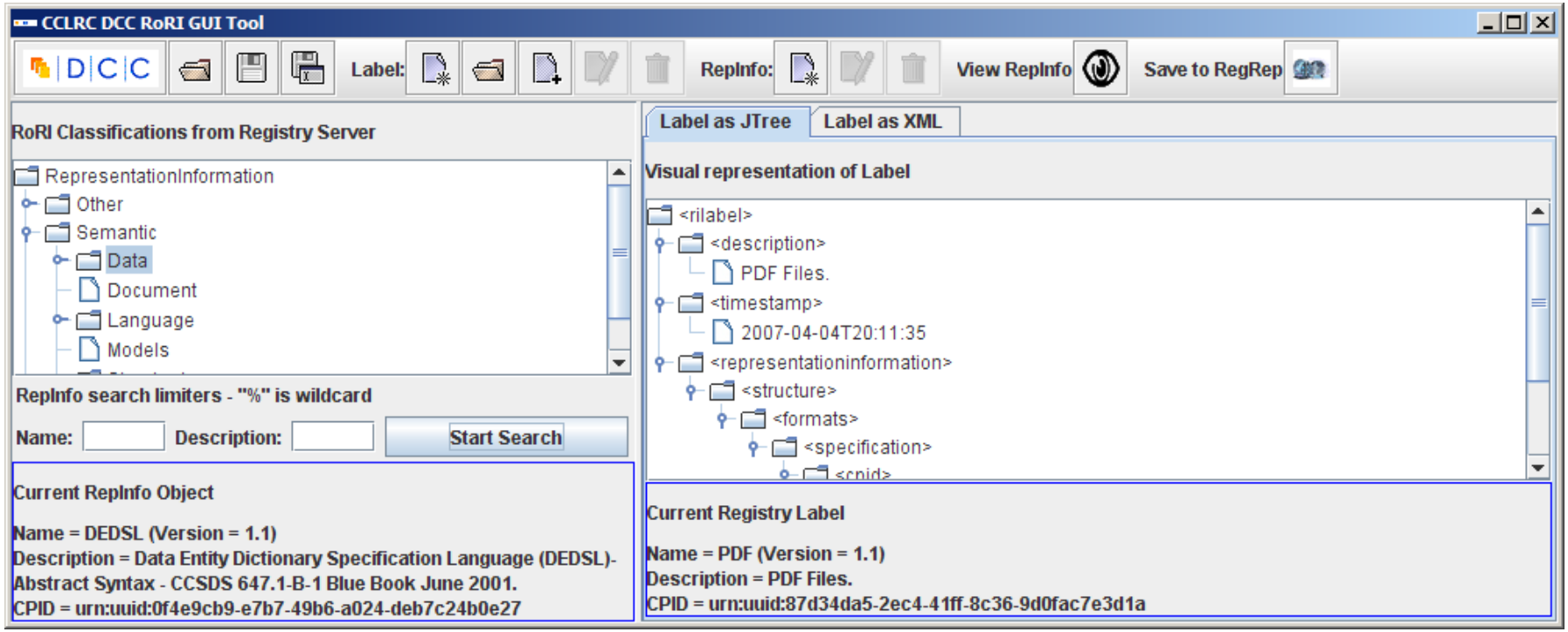
- Crystallography data are highly structured
- Convention is to share derived or reduced data, access to raw data is rare
- Crystallography Information File (CIF) is a de facto exchange standard
- CIF maintained by International Union of Crystallography (IUCr)
- Open standards and software e.g. CIF, checkcif, CML, INChI
- Culture for sharing/depositing data (CCDC)
- Well-established workflow for crystallography experiments



CIF RI Network (Partial)



RRoRI GUI Tool



A client with a GUI for ingest, search and retrieval of RI and RI labels

Conclusions

- A preservation strategy based on RI depends on a global, well-engineered, distributed infrastructure of RI
  - Needs coordination, collaboration and globally shared effort
  - Mining of RI networks for inference purposes
- Creation of robust RI networks requires domain expertise
- Likely to be gaps in global networks of RI
  - Business case for using a store of RI is clear, however the case for submitting RI to the global effort is less clear (commercial, IPR etc.)

Thank You

Questions?

Manjula Patel, Simon Coles

eCrystals Federation Project
http://wiki.ecrystals.chem.soton.ac.uk/index.php/Main_Page